

# Maximally informative “stimulus energies” in the analysis of neural responses to natural signals

Kanaka Rajan and William Bialek

*Joseph Henry Laboratories of Physics and Lewis–Sigler Institute for Integrative Genomics,  
Princeton University, Princeton, New Jersey 08544*

(Dated: January 4, 2012)

The concept of feature selectivity in sensory signal processing can be formalized as dimensionality reduction: in a stimulus space of very high dimensions, neurons respond only to variations within some smaller, relevant subspace. But if neural responses exhibit invariances, then the relevant subspace typically cannot be reached by a Euclidean projection of the original stimulus. We argue that, in several cases, we can make progress by appealing to the simplest nonlinear construction, identifying the relevant variables as quadratic forms, or “stimulus energies.” Natural examples include non–phase–locked cells in the auditory system, complex cells in visual cortex, and motion–sensitive neurons in the visual system. Generalizing the idea of maximally informative dimensions, we show that one can search for the kernels of the relevant quadratic forms by maximizing the mutual information between the stimulus energy and the arrival times of action potentials. Simple implementations of this idea successfully recover the underlying properties of model neurons even when the number of parameters in the kernel is comparable to the number of action potentials and stimuli are completely natural. We explore several generalizations that allow us to incorporate plausible structure into the kernel and thereby restrict the number of parameters. We hope that this approach will add significantly to the set of tools available for the analysis of neural responses to complex, naturalistic stimuli.

## I. INTRODUCTION

A central concept in neuroscience is feature selectivity: as our senses are bombarded by complex, dynamic inputs, individual neurons respond to specific, identifiable components of these data [1, 2]. Neurons early in a processing pathway are thought to be sensitive to simpler features [3, 4], and one can think of subsequent stages of processing as computing conjunctions of these features, so that neurons later in the pathway respond to more complex structures in the sensory world [5]. A major challenge for theory is to make this intuition mathematically precise, and to use such a precise formulation to build tools that allow us to analyze real neurons as they respond to naturalistic inputs. There is a long history of such work, but much of it rests on the identification of “features” with filters or templates. Filtering is a linear operation, and matching to a template can be thought of as a cascade of linear and nonlinear steps. As we will see, however, there are many examples of neural feature selectivity, well known from experiments on visual and auditory systems in many organisms, for which such a description in linear terms does not lead to much simplification.

In this paper we use examples to motivate the simplest nonlinear definition of a feature, in which the relevant variable is a quadratic form in stimulus space. Because the resulting variable is connected to the “energy in frequency bands” for auditory signals, we refer to these quadratic forms as “stimulus energies.” To be useful, we have to be able to identify these structures in experiments where neurons are driven by complex, naturalistic inputs. We show that, generalizing the idea of maximally informative dimensions [6], we can find the maximally

informative stimulus energies using methods that don’t require special assumptions about the structure of the input stimulus ensemble. We illustrate these ideas on model neurons, and explore the amount of data that will be needed to use these methods in the analysis of real neurons.

## II. MOTIVATION

To motivate the problems that we address, let us start by thinking about an example from the auditory system. This starting point is faithful to the history of our subject, since modern approaches for estimating receptive fields and filters have their origins in the classic work of de Boer and coworkers on the “reverse correlation” method [7], which was aimed at separating the filtering of acoustic signals by the inner ear from the nonlinearities of spike generation in primary auditory neurons. We will see that mathematically identical problems arise in thinking about complex cells in visual cortex, motion sensitive neurons throughout the visual pathway, and presumably in other problems as well.

We begin with the simplest model of an auditory neuron. If the sound pressure as a function of time is  $s(t)$ , it is plausible that the activity of a neuron is controlled by some filtered version of this stimulus, so that the probability per unit time of generating a spike is

$$r(t) = r_0 g \left[ \int d\tau f(\tau) s(t - \tau) \right], \quad (1)$$

where  $f(\tau)$  is the relevant temporal filter and  $g[\cdot]$  is a nonlinearity; the spikes occur at times  $t_i$ . The statement that neurons are tuned is that if we look at the filter in

Fourier space,

$$\tilde{f}(\omega) = \int dt f(t) e^{i\omega t}, \quad (2)$$

then the magnitude of the filter response,  $|\tilde{f}(\omega)|$ , has a relatively sharp peak near some characteristic frequency  $\omega_c$ . If we choose the stimulus waveforms from a Gaussian white noise ensemble, then the key result of reverse correlation is that if we compute the average stimulus in the neighborhood of a spike, we will recover the underlying filter, independent of the nonlinearity,

$$\langle s(t - t_i) \rangle \propto f(-t). \quad (3)$$

We emphasize that this is a theorem, not a heuristic data analysis method. *If* the conditions of the theorem are met, then this analysis is guaranteed to give the right answer in the limit of large amounts of data. If the conditions of the theorem are not met, then the spike-triggered average stimulus need not correspond to any particular characteristic of the neuron.

Eq. (1) is an example of dimensionality reduction. In principle, the neuron's response at time  $t$  can be determined by the entire history of the stimulus for times  $t' \leq t$ . Let us suppose that we sample (and generate) the stimulus in discrete time steps spaced by  $dt$ . Then the stimulus history is a list of numbers

$$\mathbf{s}_t \equiv \{s(t), s(t - dt), s(t - 2dt), \dots, s(t - Ddt)\}, \quad (4)$$

where  $D$  is the effective stimulus dimensionality, set by  $D = T/dt$ , and  $T$  the longest plausible estimate of the integration time for the neural response. We can think of  $\mathbf{s}_t$  as a  $D$ -dimensional vector. If we know that the neural response is controlled by a linearly filtered version of the sound pressure stimulus, even followed by an arbitrary nonlinearity, then only one direction in this  $D$ -dimensional space matters for the neuron. Further this really is a “direction,” since we can write the response as the Euclidean projection of  $\mathbf{s}$  onto one axis, or equivalently the dot product between  $\mathbf{s}$  and a vector  $\mathbf{W}$ ,

$$r(t) = r_0 g(\mathbf{W} \cdot \mathbf{s}_t), \quad (5)$$

where

$$\mathbf{W} = dt \times \{f(0), f(dt), f(2dt), \dots, f(T)\}. \quad (6)$$

This explicit formulation in terms of dimensionality reduction suggests a natural generalization in which several dimensions, rather than just one, are relevant,

$$r(t) = r_0 g(\mathbf{W}_1 \cdot \mathbf{s}_t, \mathbf{W}_2 \cdot \mathbf{s}_t, \dots, \mathbf{W}_K \cdot \mathbf{s}_t). \quad (7)$$

As long as we have  $K \ll D$ , it still holds true that the neuron responds only to some limited set of stimulus dimensions, but this number is not as small as in the simplest model of a single filter.

Notice that if an auditory neuron responds according to Eq (1), then it will exhibit “phase locking” to

periodic stimuli. Specifically, if  $s(t) = A \cos(\omega t)$  and  $\tilde{f}(\omega) = |\tilde{f}(\omega)| e^{+i\phi}$ , then  $r(t) = r_0 g[A|\tilde{f}(\omega)| \cos(\omega t - \phi)]$ . So long as there is a nonzero response to the stimulus, this response will be modulated at the stimulus frequency  $\omega$ , and more generally if we plot the spiking probability versus time measured by the phase  $\psi = \omega t$  of the stimulus oscillation, then the probability will vary with, or “lock” to this phase.

While almost all auditory neurons are tuned, not all exhibit phase locking. We often summarize the behavior of tuned, non-phase-locked neurons by saying that they respond to the power in a given bandwidth or to the envelope of the signal at the output of a filter. The simplest model for such behavior, which has its roots in our understanding of hair cell responses [8–11], is to imagine that the output of a linear filter passes through a weak nonlinearity, then another filter. The second stage of filtering is low-pass, and will strongly attenuate any signals at or near the characteristic frequency  $\omega_c$ . Then, to lowest order, the neuron's response depends on

$$p(t) = \int d\tau f_2(\tau) \left[ \int dt' f_1(t - \tau - t') s(t') \right]^2, \quad (8)$$

where  $f_1$  is the bandpass filter that determines the tuning of the neuron and  $f_2$  is a smoothing filter which ensures that the cell responds to the power in its preferred frequency band while rejecting any temporal variation at or above the “carrier” frequency. The probability of spiking depends on this power  $p(t)$  through a nonlinearity, as before,

$$r(t) = r_0 g[p(t)]. \quad (9)$$

Intuitively, this simple model for a non-phase-locked neuron also represents a substantial reduction in dimensionality – all that matters is the power passing through a given frequency band, defined by the filter  $f_1$ . On the other hand, we cannot collapse this model into the one dimensional form of Eq. (5). To be concrete, suppose that the filter  $f_1$  has a relatively narrow bandwidth around its characteristic frequency  $\omega_c$ . Then we can write

$$f_1(\tau) = A(\tau) \sin(\omega_c \tau + \phi), \quad (10)$$

where the amplitude  $A(\tau)$  varies slowly compared to the period of the oscillation. Let us denote by  $\tau_1$  the temporal width of  $A(\tau)$ , since this corresponds to the time over which the filter  $f_1$  integrates the stimulus, and similarly  $\tau_2$  will denote the temporal width of  $f_2(\tau)$ . To make sure that the power  $p(t)$  does not oscillate at (twice) the frequency  $\omega_c$ , we must force  $\tau_2 \gg 2\pi/\omega_c$ . But we still have two possibilities, (a)  $\tau_1 \gg \tau_2 \gg \omega_c$  and (b)  $\tau_2 \gg \tau_1 \gg \omega_c$ . If (a) is true, we can show that  $p(t)$  is the Pythagorean sum of the outputs of two filters that form a quadrature pair,

$$p(t) \approx \left[ \int dt' f_\alpha(t - \tau - t') s(t') \right]^2$$

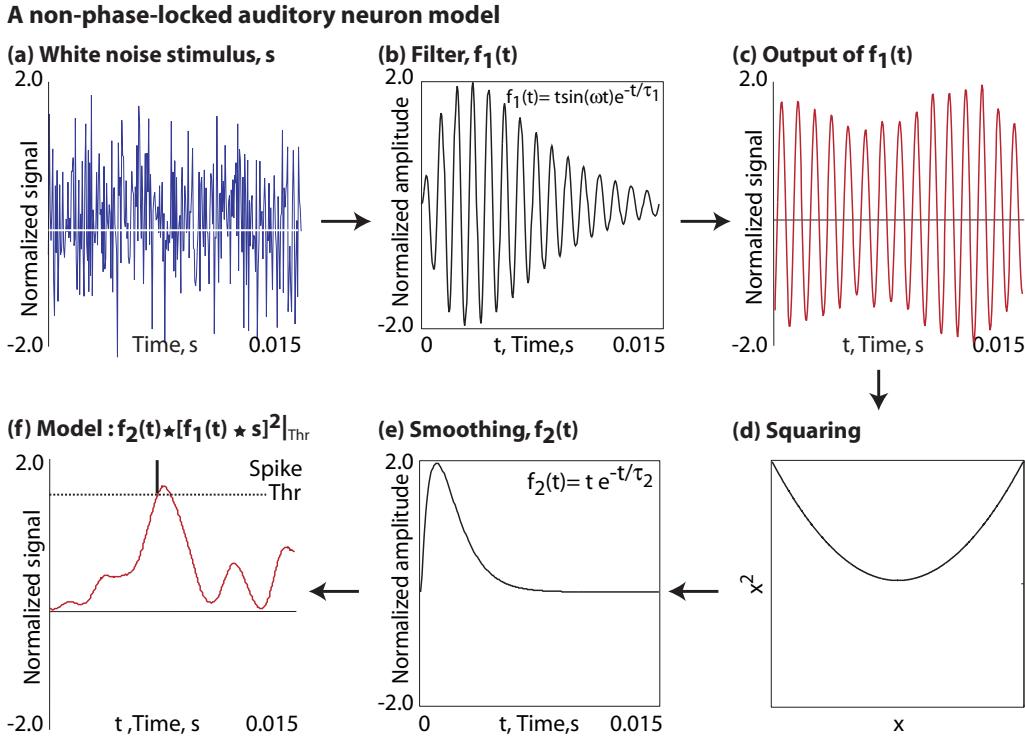


FIG. 1. **A non-phase-locked auditory neuron.** (a) In this implementation a model neuron responds to a white noise stimulus. (b) The stimulus  $s$  is filtered through a temporal filter  $f_1$ , which has the form  $t \sin(\omega t) \exp(-t/\tau_1)$  where  $\omega = 2\pi \times 10^3$  Hz and  $\tau_1 = 3$  ms. (c) The output of  $f_1$  is shown here. The filter  $f_1$  is narrow band, therefore the output oscillates at the characteristic frequency even when the input is white. (d) The output of  $f_1$  is first squared, and in (e), convolved with a second filter  $f_2$  of the form  $t \exp(-t/\tau_2)$ , with a smoothing time constant  $\tau_2 = 1$  ms. (f) The normalized signal is finally thresholded to generate spikes at a mean rate that allows the spiking probability to be 0.1 per time step. We assume that time runs discretely in steps of  $dt = 1/20000$  s.

$$+ \left[ \int dt' f_\beta(t - \tau - t') s(t') \right]^2, \quad (11)$$

where

$$f_\alpha(\tau) \approx A(\tau) \sin(\omega_c \tau + \phi) \quad (12)$$

$$f_\beta(\tau) \approx A(\tau) \cos(\omega_c \tau + \phi). \quad (13)$$

On the other hand, if (b) is true, there is no simple decomposition, and the minimum number of dimensions that we need to describe this model is  $K \sim \tau_2/dt$ , which can be quite large.

We can measure the number of relevant dimensions using the spike-triggered covariance matrix. Specifically, if the stimulus vector  $\mathbf{s}_t$  has components  $s_i(t)$ , then we can form the matrix

$$\Delta C_{ij} = \langle s_i(t) s_j(t) \rangle_{t=t_{\text{spike}}} - \langle s_i(t) s_j(t) \rangle, \quad (14)$$

where in the first term we average over the arrival time of the spikes and in the second term we average over all time. If the spiking probability behaves as in Eq. (7), and we choose the stimulus from a Gaussian ensemble, then  $\Delta C$  has exactly  $K$  nonzero eigenvalues [12, 13]. In Fig. 1

we schematize the model auditory neuron we have been describing, and in Fig. 2 we show the spike-triggered covariance analysis for models in the two limits,  $\tau_2 \ll \tau_1$  and  $\tau_2 \gg \tau_1$ . Indeed we find that in the first case there are just two relevant dimensions, a quadrature pair of filters, whereas in the second case there are many relevant dimensions; these dimensions appear as temporally shifted and orthogonalized copies of the filter  $f_1(t)$ .

We can think of a neuron that does not phase lock as having an invariance: it responds to acoustic waveforms that have energy in a relevant bandwidth near  $\omega_c$ , but it doesn't discriminate among signals that are shifted by small times. This invariance means that the cell is not just sensitive to one dimension of the stimulus, but to many, although these different dimensions correspond, in effect, to the same stimulus feature occurring at different times relative to the spike. Thus, we have a conflict between the notion of a “single feature” and the mathematical description of a “single dimension” via linear projection. The challenge is to provide a mathematical formulation that better captures our intuition. Before presenting a possible solution, let's see how the

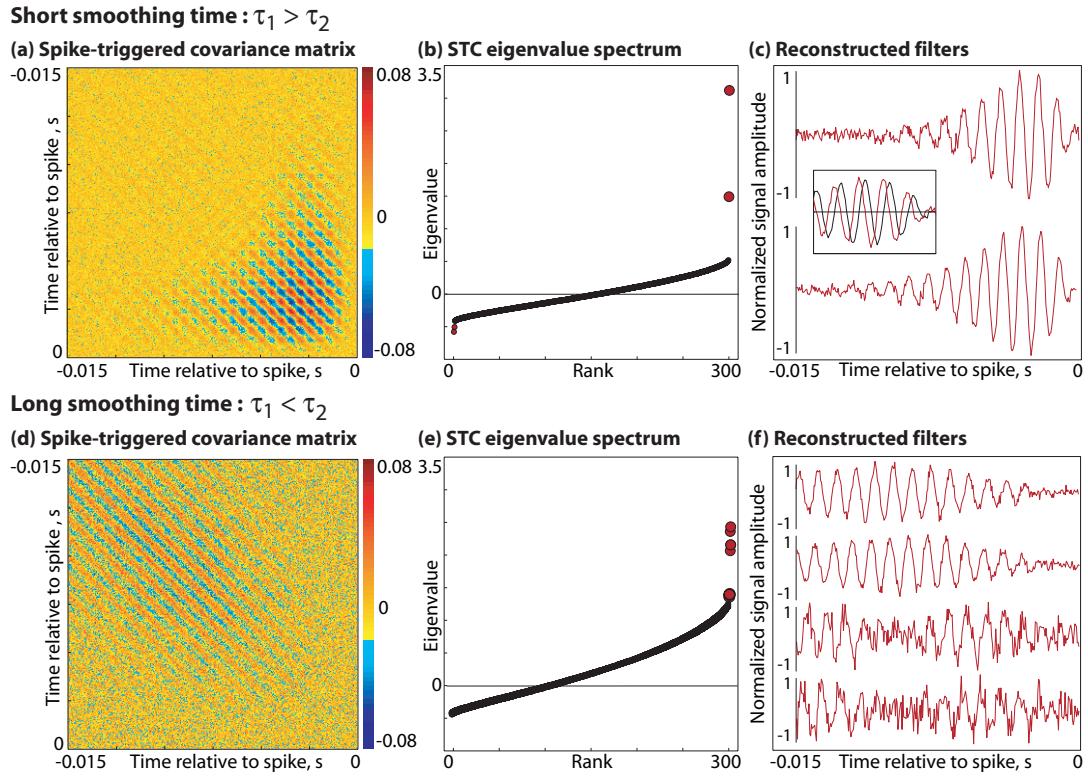


FIG. 2. Covariance analysis of the non-phase-locked auditory neuron. (a) If the time constant of the smoothing filter  $f_2$  is much shorter than that of the filter  $f_1$ , the spike-triggered covariance matrix has a relatively simple structure. (b) Diagonalizing this covariance matrix yields 2 leading eigenvalues (the rest remain close to 0). (c) The eigenvectors corresponding to the 2 non-zero eigenvalues are the reconstructed filters plotted here. These form a quadrature pair, as shown in the inset. (d) If the smoothing time of the filter  $f_2$  is larger than that of the first, the covariance matrix has a much richer structure. (e) The spike-triggered covariance matrix decomposes into multiple non-unique eigenvalues. (f) The eigenvectors corresponding to the non-zero eigenvalues give multiple time-shifted copies of the same filter.

same problem arises in other cases.

Since the classical work of Hubel and Wiesel [14, 15], we know that cells in the primary visual cortex can be classified as “simple” and “complex.” Although Hubel and Wiesel did not give a mathematical description of their data, in subsequent work, simple cells often have been described in the same way that we described the simplest auditory neuron in Eq (1) [16]. If the light intensity falling on the retina varies in space ( $\vec{x}$ ) and time ( $t$ ) as  $I(\vec{x}, t)$ , we can define a spatiotemporal receptive field  $F(\vec{x}, \tau)$  and approximate the probability that a simple cell generates a spike per unit time as

$$r(t) = r_0 g \left[ \int d^2x \int d\tau F(\vec{x}, \tau) I(\vec{x}, t) \right]. \quad (15)$$

If, as before, we assume that the stimulus is generated in discrete time steps (movie frames) with spacing  $dt$ , and that the stimulus influences spikes only within some time window of duration  $T$ , then we can think of the stimulus at any moment in time as being the  $T/dt$  frames of the

movie preceding that moment,

$$\mathbf{s}_t \equiv \{I(\vec{x}, t), I(\vec{x}, t - dt), I(\vec{x}, t - 2dt), \dots, I(\vec{x}, t - T)\}. \quad (16)$$

If the relevant region of space is within  $d \times d$  pixels, then this stimulus vector lives in a space of dimension  $D = d^2 T/(dt)$ , which can be enormous. As in the discussion above, Eq. (15) is a restatement of the hypothesis that only one direction in this space is relevant for determining the probability that the simple cell generates a spike, and “direction” is once again a Euclidean or linear projection.

For a complex cell, on the other hand, this single projection is inadequate. Complex cells respond primarily to oriented edges and gratings, as do simple cells, but they have a degree of spatial invariance which means that their receptive fields cannot be mapped onto fixed zones of excitation and inhibition. Instead, they respond to patterns of light in a certain orientation within a large receptive field, regardless of precise location, or to movement in a certain direction. Corresponding to this intuition, analysis of complex cells using the spike-triggered covariance method shows that there is more than one relevant dimension [17]. As with non-phase-locked au-

ditory neurons, what defeats the simplest version of dimensionality reduction in complex cells is the invariance of the response, in this case, invariance to small spatial displacement of the relevant, oriented stimulus feature.

The simplest model of a complex cell is precisely analogous to the quadrature pair of filters that emerge in the analysis of non-phase-locked auditory neurons. To be concrete, let us imagine that receptive fields are described by Gabor patches. The position  $\vec{x}$  includes two orthogonal coordinates in visual space, which we call  $x_1$  and  $x_2$ . Gabor patches have an oscillatory dependence on one of these coordinates, but simply integrate along the other; both the integration and the envelope of the oscillations are described by Gaussians, so that

$$F_1(\vec{x}, \tau) = \cos(kx_1) \exp\left[-\left(\frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}\right)\right] f(\tau), \quad (17)$$

and the quadrature filter is then

$$F_2(\vec{x}, \tau) = \sin(kx_1) \exp\left[-\left(\frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}\right)\right] f(\tau). \quad (18)$$

Each of these filters is maximally sensitive to extended features oriented along the  $x_2$  direction, but the optimal

patterns of light and dark are shifted for the two filters; for simplicity we have assumed that spatial and temporal filtering is separable. If we form the energy-like quantity

$$p(t) = \left[ \int d^2x \int d\tau F_1(\vec{x}, \tau) I(\vec{x}, t) \right]^2 + \left[ \int d^2x \int d\tau F_2(\vec{x}, \tau) I(\vec{x}, t) \right]^2, \quad (19)$$

we have a measure of response to oriented features independent of their precise position, and this provides a starting point for the analysis of a complex cell.

One more example of the problem we face is provided by the computation of motion in the visual system. There is a classical model for this computation, the correlator model, that grew out of the experiments by Hassenstein and Reichardt on behavioral responses to visual motion in beetles and flies [19]. Briefly, the idea behind this model is that if something is moving at velocity  $v$ , then the image intensity  $I(x, t)$  must be correlated with the intensity at  $I(x + \Delta, t + \tau)$ , where  $\tau = \Delta/v$ . Then we can detect motion by computing this correlation and averaging over some window in space and time,

$$\mathbf{C}_{\Delta, \tau}(t) = \int dx W(x) \int dt' f(t - t') I(x + \Delta, t') I(x + \Delta, t' + \tau), \quad (20)$$

where for simplicity, we think just about a single spatial dimension. In principle, with just one value of the delay  $\tau$  and one value of the displacement  $\Delta$ , this correlation “detects” only motion at one velocity,  $v = \Delta/\tau$ , but we can easily generalize this computation to a weighted sum over different values of these spatiotemporal shifts,

$$\mathbf{C}(t) = \int dx W(x) \int dt' f(t - t') \int d\tau \sum_{\Delta} F(\Delta, \tau) I(x + \Delta, t') I(x + \Delta, t' + \tau). \quad (21)$$

Depending on the precise form of this weighting we can arrange for the correlation  $\mathbf{C}$  to have a relatively smooth, graded dependence on the velocity.

In the insect visual system it seems natural to think of the correlations in Eq. (21) as being computed from the outputs of individual photoreceptors, which are typically spaced  $\sim 1^\circ$  apart. In mammals, we can imagine computing a similar quantity, but we would use the outputs of larger retinal or cortical receptive fields [18]. We can also think about this computation in Fourier space. If we transform

$$I(x, t) = \int \frac{dk}{2\pi} \int \frac{d\omega}{2\pi} \tilde{I}(k, \omega) e^{ikx - i\omega t}, \quad (22)$$

then we can think of  $\mathbf{C}$  as integrating power or energy  $P(k, \omega) = |\tilde{I}(k, \omega)|^2$  over some region in the  $k - \omega$  plane; motion corresponds to having this power concentrated along the line  $\omega = vk$ . Once again there is an invariance to this computation, since as with the non-phase-locked auditory neuron we are looking for power regardless of

phase. More directly, in this case the brain is trying to compute a velocity, more or less independently of the absolute position of the moving objects. Even in an insect visual system, where computing  $\mathbf{C}$  corresponds to correlating the filtered outputs of neighboring photoreceptors in the compound eye, this computation is repeated across an area containing many photoreceptors, and hence there is no way to collapse  $\mathbf{C}$  down to a function of just one or two Euclidean projections in the stimulus space.

What do these examples – non-phase-locked auditory neurons, complex cells in the visual cortex, and visual motion detectors – have in common? In all three cases, the natural, simplest starting point is a model in which the brain computes not a linear projection of the stimulus onto a receptive field, but rather a quadratic form. More precisely, if stimuli are the vectors  $\mathbf{s} = \{s_1, s_2, \dots, s_D\}$ , then Eq’s. (8), (19), and (21) all correspond to comput-

ing a quantity

$$x = \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s} \equiv \sum_{i,j=1}^D s_i Q_{ij} s_j. \quad (23)$$

Generalizing the use of terms such as “energy in a frequency band” and “motion energy,” we refer to  $x$  as a “stimulus energy.”

If the matrix  $\mathbf{Q}$  is of low rank, this means that we can build  $x$  out of projections of the stimulus onto a correspondingly low dimensional Euclidean subspace, and we can try to recover the relevant structure using methods such as the spike-triggered covariance or maximally informative dimensions. Still, if  $\mathbf{Q}$  is of rank 5 (for example), once we identify the five relevant dimensions we have to explore the nonlinear input/output relation of the neuron thoroughly enough to recognize whether these dimensions combine in a simple way, to recover our intuition that there is a single stimulus feature  $x$  to which the cell responds. This can be prohibitively difficult.

In many cases, it is plausible that neurons could be responding just to one energy  $x$ , but the underlying matrix  $\mathbf{Q}$  could be of full rank; a clear example is provided by correlator models for wide-field motion sensitive neurons, as in [12, 19]. But in this case there is no real “dimensionality reduction” associated with the mapping from  $\mathbf{s} \rightarrow x$ , if all we know how to do is to search for linear projections or Euclidean subspaces. On the other hand, mapping  $\mathbf{s} \rightarrow x$  really is a tremendous reduction in complexity, because the full stimulus  $\mathbf{s}$  is described by  $D$  parameters, while  $x$  is just one number.

Suppose that the response of a neuron to complex stimuli can be described by saying that the probability of spiking depends solely on a single stimulus energy  $x$  as in Eq. (23), so that

$$r(t) = r_0 g(\mathbf{s}_t^T \cdot \mathbf{Q} \cdot \mathbf{s}_t). \quad (24)$$

Our task becomes one of showing that we can recover the underlying matrix  $\mathbf{Q}$  by analyzing the spike train in relation to the stimuli, without making any assumptions about the statistics of the stimuli.

As we were finishing this work, we became aware of two other recent efforts that point to the importance of quadratic forms in stimulus space. It is shown in [27] that a logistic dependence of the spike probability on a quadratic form is the maximum entropy, and hence least structured, model consistent with a measurement of the spike-triggered covariance matrix, and that this identification is correct without any assumptions about the distributions of inputs. As a practical matter, Ref [27] emphasizes the case where the matrix kernel (corresponding to  $\mathbf{Q}$  above) is of low rank, and explores examples in which the quadratic form provides an incremental improvement on linear dimensionality reduction. In a different direction, the work in [28] consider models in which the spike probability depends exponentially on a quadratic form, and spiking is explicitly a

Poisson process. They show that this model, and some generalizations, lends itself to a Bayesian formulation, in which various simplifying structures (see Section IV) can be imposed through prior distributions on the possible  $\mathbf{Q}$ , although it is not clear whether computationally tractable priors correspond to realistic models. In some limits, these approaches are equivalent to one another, and to the search for maximally informative stimulus energies that we propose here. As far as we can see, the Maximally Informative Stimulus Energy method always involves fewer assumptions, and can capture any of the structures postulated in the other methods.

### III. CORE OF THE METHOD

If the probability of generating an action potential depends on the stimulus  $\mathbf{s}$ , then observing the arrival of even a single spike provides information about the stimulus. Importantly, the data processing inequality [25] tells us that if we look not at the full stimulus but only at some limited or compressed description of the stimulus – a single feature, for example – we can only lose information. If the neuron really is sensitive to only one stimulus feature, however, it is possible to lose none of the available mutual information between spikes and stimuli by focusing on this one feature. This suggests a natural strategy for identifying relevant stimulus features, by searching for those which preserve the maximum amount of information [6]. Further, we can put the success of such a search on an absolute scale, by estimating more directly the information that spikes provide about the stimulus [20, 21], and asking what fraction of this information is captured by the best feature we could find.

#### A. Setting up the problem

To make this precise, we recall that the information about the stimulus  $\mathbf{s}$  that is conveyed by the observation of a single spike can be written, following [20], as

$$I_{\text{spike}} = \int d^D \mathbf{s} P(\mathbf{s}|\text{spike}) \log_2 \left[ \frac{P(\mathbf{s}|\text{spike})}{P(\mathbf{s})} \right] \text{ bits}, \quad (25)$$

where  $P(\mathbf{s})$  is the stimulus distribution and  $P(\mathbf{s}|\text{spike})$  is the distribution of stimuli given that a spike occurred at a particular time (the response conditional ensemble [22]). If we consider a mapping  $\mathcal{M} : \mathbf{s} \rightarrow x$ , then we can also compute

$$I_{\text{spike}}(\mathcal{M}) = \int dx P(x|\text{spike}) \log_2 \left[ \frac{P(x|\text{spike})}{P(x)} \right], \quad (26)$$

knowing that for any mapping  $\mathcal{M}$ ,  $I_{\text{spike}}(\mathcal{M}) \leq I_{\text{spike}}$ . If we restrict ourselves to some class of mappings, then we can search for an  $\mathcal{M}^*$  which maximizes  $I_{\text{spike}}(\mathcal{M}^*)$ , and see how close this is to the real value of  $I_{\text{spike}}$ . If the inputs are Gaussian, then we can find  $\mathcal{M}$  by standard

spike-triggered or reverse correlation methods. Working with arbitrary (i.e., natural) stimulus ensembles complicates things. Further, if there is just one stimulus feature  $x$ , then computing  $I_{\text{spike}}(\mathcal{M})$  involves only probability distributions over a single variable, so there is no curse of dimensionality. Finally, in order for this to work, we need to make no assumptions about the form of the distribution of stimuli  $P(\mathbf{s})$ , or the inherited distribution of stimulus features  $P(x)$ . Thus, as first emphasized in [6], searching for maximally informative features provides a practical and principled way to analyze neural responses to high dimensional, naturalistic stimuli.

The work in [6] considered the case where the stimulus features are one or more linear projections of the stimulus,  $x_\alpha = \mathbf{W}_\alpha \cdot \mathbf{s}$ , so that the mapping  $\mathcal{M}$  is parameterized by the vectors  $\{\mathbf{W}_\alpha\}$ ; in the simplest case there being just one vector  $\mathbf{W}$ . Here we are interested in quadratic mappings, corresponding to the stimulus energies in Eq. (23). Now the mapping  $\mathcal{M}$  is parameterized by a symmetric matrix  $\mathbf{Q}$ . In principle, all the arguments of [6] for the linear case should generalize to the quadratic case, and we follow this path. Before starting, we note an obvious problem related to the number of parameters we are looking for. If we are searching for a vector  $\mathbf{W}$  in a  $D$ -dimensional stimulus space, we are looking an object described by  $D$  numbers. In fact, the length of the vector is irrelevant, so that maximizing  $I_{\text{spike}}(\mathcal{M})$  corresponds to optimization in a  $D - 1$  dimensional space. But if we are searching for a symmetric matrix that acts on stim-

ulus vectors, there are  $D(D + 1)/2 - 1$  free parameters. This is a problem both because we have to optimize in a space of much higher dimensionality, and because determining more parameters reliably must require larger data sets. We will address these problems shortly, but let's start by following the path laid out in [6], which involves searching for the maximum of  $I_{\text{spike}}(\mathcal{M})$  by gradient ascent.

If our stimulus feature is the energy in Eq. (23), then the distribution of  $x$  is

$$P_{\mathbf{Q}}(x) = \int d^D s \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}), \quad (27)$$

where the subscript explicitly denotes that  $P(x)$  depends on  $\mathbf{Q}$ . We take the derivative of this with respect to an element  $Q_{ij}$  in the matrix  $\mathbf{Q}$ ,

$$\frac{\partial P_{\mathbf{Q}}(x)}{\partial Q_{ij}} = \int d^D s \frac{\partial}{\partial Q_{ij}} \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) \quad (28)$$

$$= - \int d^D s s_i s_j \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}). \quad (29)$$

Similarly, we can differentiate the distribution of  $x$  conditional on a spike,

$$\frac{\partial P_{\mathbf{Q}}(x|\text{spike})}{\partial Q_{ij}} = - \int d^D s s_i s_j \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}). \quad (30)$$

Putting these terms together, we can differentiate the information:

$$\frac{\partial I_{\text{spike}}(\mathcal{M})}{\partial Q_{ij}} = \int dx \left[ \frac{\delta I_{\text{spike}}(\mathcal{M})}{\delta P_{\mathbf{Q}}(x)} \frac{\partial P_{\mathbf{Q}}(x)}{\partial Q_{ij}} + \frac{\delta I_{\text{spike}}(\mathcal{M})}{\delta P_{\mathbf{Q}}(x|\text{spike})} \frac{\partial P_{\mathbf{Q}}(x|\text{spike})}{\partial Q_{ij}} \right] \quad (31)$$

$$= \frac{1}{\ln 2} \int dx \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \int d^D s s_i s_j \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) - \frac{1}{\ln 2} \int dx \left( 1 + \ln \left[ \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \right] \right) \int d^D s s_i s_j \delta'(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}) \quad (32)$$

$$= - \frac{1}{\ln 2} \int dx \int d^D s s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \right] + \frac{1}{\ln 2} \int dx \frac{P_{\mathbf{Q}}(x)}{P_{\mathbf{Q}}(x|\text{spike})} \int d^D s s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}) \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \right]. \quad (33)$$

But we notice that

$$\int d^D s s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}) = \langle s_i s_j | x \rangle P_{\mathbf{Q}}(x), \quad (34)$$

where  $\langle s_i s_j | x \rangle$  is the expectation value of  $s_i s_j$  conditional on the value of the stimulus energy  $x$ , and similarly

$$\int d^D s s_i s_j \delta(x - \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}) P(\mathbf{s}|\text{spike}) = \langle s_i s_j | x, \text{spike} \rangle P_{\mathbf{Q}}(x|\text{spike}), \quad (35)$$

where  $\langle s_i s_j | x, \text{spike} \rangle$  is the expectation value conditional on the energy  $x$  and the occurrence of a spike. We can combine these terms to give

$$\frac{\partial I_{\text{spike}}(\mathcal{M})}{\partial Q_{ij}} = \int dx P_{\mathbf{Q}}(x) [\langle s_i s_j | x, \text{spike} \rangle - \langle s_i s_j | x \rangle] \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \right], \quad (36)$$

or, more compactly,

$$\nabla_{\mathbf{Q}} I = \int dx P_{\mathbf{Q}}(x) [\langle \mathbf{s} \mathbf{s}^T | x, \text{spike} \rangle - \langle \mathbf{s} \mathbf{s}^T | x \rangle] \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}}(x|\text{spike})}{P_{\mathbf{Q}}(x)} \right]. \quad (37)$$

To learn the maximally informative energy, or the best choice of the matrix  $\mathbf{Q}$ , we can ascend the gradient in successive learning steps,

$$\mathbf{Q} \rightarrow \mathbf{Q} + \gamma \nabla_{\mathbf{Q}} I \text{ where } \gamma \text{ is small.} \quad (38)$$

## B. Multiple matrices

In the same way that the idea of linear projection can be generalized to have the probability of spiking depend on multiple linear projections, we can generalize to the case where there are multiple relevant stimulus energies. Perhaps the simplest example Eq. (24) can be generalized to 2 matrices, is the computation of a (regularized) ratio between two stimulus energies, so that the probability of spiking varies as

$$r = r_0 g \left( \frac{\mathbf{s}^T \cdot \mathbf{Q}_1 \cdot \mathbf{s}}{1 + \mathbf{s}^T \cdot \mathbf{Q}_2 \cdot \mathbf{s}} \right). \quad (39)$$

Some biological examples of this formulation include gain control or normalization in V1 [17], optimal estimation theory of motion detection in visual neurons of insects [12] and complex spectrotemporal receptive fields of neurons responsible for song processing in songbirds [23, 24]. The inference task becomes one of estimating both matrices  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$  by information maximization.

As before, we can compute the gradient, noting that this time, there are two different gradients of  $I_{\text{spike}}(\mathcal{M})$ ,

$$\begin{aligned} \nabla_{\mathbf{Q}_1} I &= \int dx P_{\mathbf{Q}_1}(x) \left[ \left\langle \frac{\mathbf{s} \mathbf{s}^T}{1 + \mathbf{s}^T \cdot \mathbf{Q}_2 \cdot \mathbf{s}} \middle| x, \text{spike} \right\rangle - \left\langle \frac{\mathbf{s} \mathbf{s}^T}{1 + \mathbf{s}^T \cdot \mathbf{Q}_2 \cdot \mathbf{s}} \middle| x \right\rangle \right] \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}_1}(x|\text{spike})}{P_{\mathbf{Q}_1}(x)} \right] \text{ and,} \\ \nabla_{\mathbf{Q}_2} I &= - \int dx P_{\mathbf{Q}_2}(x) \left[ \left\langle \frac{\mathbf{s}^T \cdot \mathbf{Q}_1 \cdot \mathbf{s}}{(1 + \mathbf{s}^T \cdot \mathbf{Q}_2 \cdot \mathbf{s})^2} \mathbf{s} \mathbf{s}^T \middle| x, \text{spike} \right\rangle - \left\langle \frac{\mathbf{s}^T \cdot \mathbf{Q}_1 \cdot \mathbf{s}}{(1 + \mathbf{s}^T \cdot \mathbf{Q}_2 \cdot \mathbf{s})^2} \mathbf{s} \mathbf{s}^T \middle| x \right\rangle \right] \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}_2}(x|\text{spike})}{P_{\mathbf{Q}_2}(x)} \right] \end{aligned} \quad (40)$$

Analogous to Eq. (38), at every learning step, we update each matrix  $\mathbf{Q}_i$  by the appropriate  $i^{\text{th}}$  gradient,

$$\mathbf{Q}_i \rightarrow \mathbf{Q}_i + \gamma \nabla_{\mathbf{Q}_i} I. \quad (41)$$

where  $i = 1, 2$  for the  $x$  in Eq. (39). In principle the formalism in Eq. (39) could yield a more complete description of a neuron's nonlinear response properties compared to a single quadratic kernel convolving the stimulus, but there are some data-requirement challenges which we will address later.

## C. Technical aspects of optimization

In order to implement Eq. (38) as an algorithm, we have to evaluate all the relevant probability distributions and integrals. In practice, this means computing  $x$  for all stimuli, choosing an appropriate binning along the  $x$ -axis, and sampling the binned versions of the spike-triggered and prior distributions. We compute the expectation values  $\langle \mathbf{s} \mathbf{s}^T \rangle$  separately for each bin, approximate the integrals as sums over the bins, and derivatives as differences between neighboring bins. To deal with local extrema in the objective function, we use a large starting value of  $\gamma$  and gradually decrease  $\gamma$  during learning. This basic prescription can be made more sophisticated, but we do not report these technical improvements here.

An example of these ideas is shown in Fig. 3. We used very small patches from natural images as inputs, reshaping the intensities in nearby pixels into a  $D$ -component stimulus vector  $\mathbf{s}$  where  $D = 10$ . To describe the neuron we chose a random symmetric matrix  $\mathbf{K}$  to use as the kernel, and generated spikes when the stimulus energy,  $\mathbf{s}^T \cdot \mathbf{K} \cdot \mathbf{s}$ , crossed a threshold, as illustrated in Fig. 3a. We fixed the mean rate of the spiking response such that the probability of a spike occurring in one bin of duration  $dt$  is 0.1, and we generated  $\sim 1000$  spikes. We then tried to extract the neuron's receptive field by starting with a random initial matrix  $\mathbf{Q}$ , and following the gradient of mutual information, as in Eq. (38).

We let the one parameter of the algorithm,  $\gamma$ , gradually decrease from a starting value of 0.5 to 0.05, in order to minimize the fluctuations around the true maximum

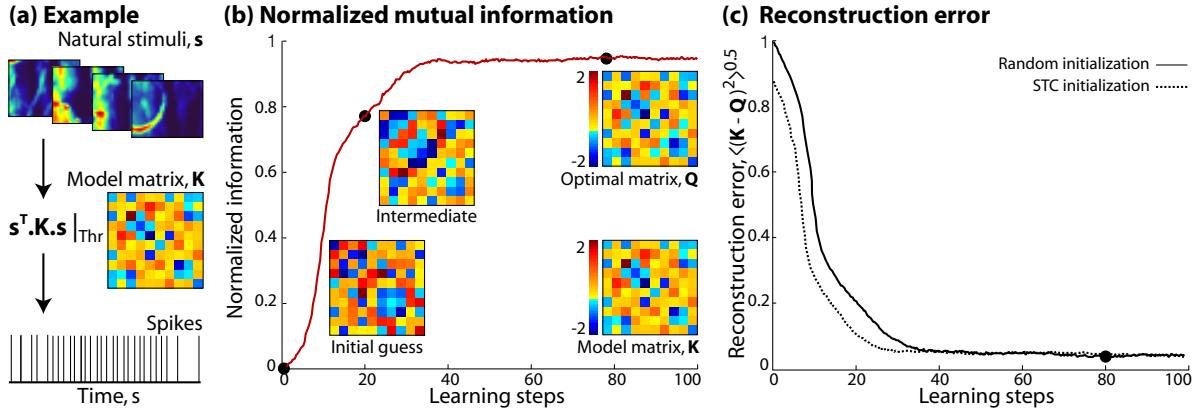


FIG. 3. **Core of the method.** (a) A general implementation is shown here. The stimuli  $s$  are natural image clips which are  $D \times D$  pixel patches resized from a natural image database, as described in [26]. 1000 spikes are generated with a probability per time bin of 0.1 from the model neuron by a thresholding the term,  $s^T \cdot K \cdot s$  where the  $10 \times 10$  matrix  $K$  is the receptive field of the neuron. (b) Mutual information between the spiking response of the model neuron and the quadratic stimulus projection  $x$  is plotted as a function of the number of learning steps. Information, normalized by its value when  $K = Q$ , peaks at the 40<sup>th</sup> learning step and then plateaus. The 3 black dots on the trace denote the points at which we extract the initial, the intermediate and the optimal matrices. The maximally informative matrix  $Q$  reconstructed at the 40<sup>th</sup> step, agrees well with  $K$ , indicating convergence. For this implementation the step size  $\gamma = 0.5$  at the start and 0.05 at the end of the algorithm. (c) Root-mean-square (RMS) reconstruction error calculated as  $\langle (K - Q)^2 \rangle^{1/2}$ , is plotted as a function of the number of learning steps. This error decreases steadily until either the randomly initialized matrix (solid line) or the matrix initialized to the spike-triggered covariance matrix (dashed line) matches  $K$ . If  $Q$  is initialized to the covariance matrix, the initial RMS error is smaller and the convergence is faster (30<sup>th</sup> learning step) compared to that for a randomly initialized  $Q$ . For this example, both  $K$  and  $Q$  are  $10 \times 10$  matrices and the black dot on the solid trace is at the same learning step as in panel (b).

of the information. Mutual information, the red trace in Fig. 3b, peaks at the 40<sup>th</sup> learning step and remains unchanged after that. The 3 black dots in Fig. 3b correspond to the steps during the optimization when we extract and plot the initial guess, the intermediate and the optimal/maximally informative matrix  $Q$ . It is interesting to note that the intermediate matrix appears completely different from the optimal  $Q$  even though the corresponding mutual information is relatively close to its maximum (a similar observation was made in the context of maximally informative dimensions [6]).

In Fig. 3c the root-mean-square (RMS) reconstruction error  $\langle (K - Q)^2 \rangle^{1/2}$  is plotted as a function of the number of learning steps for a randomly initialized  $Q$  (solid line) and when  $Q$  is initialized to the spike-triggered covariance (STC) matrix (dashed line). RMS error at the start of the algorithm  $\approx 1$  when the “true” matrix  $K$  and the initial guess for  $Q$  are symmetric, random matrices, uncorrelated with each other, but is slightly lower when  $Q$  is initialized to the STC. This difference becomes smaller as the stimulus dimensionality  $D$  increases or as the stimulus departs more strongly from Gaussianity. Both traces decrease, and stop changing once our estimate of the optimal  $Q$  matches  $K$ . This occurs at the 40<sup>th</sup> step for the randomly initialized  $Q$  and slightly sooner (30<sup>th</sup> step) when initialized to the STC. If we had fewer spikes, our estimate for the optimal  $Q$  could still match  $K$  adequately, but the actual RMS error in reconstruction might be higher. We explore such performance

measures and data requirement issues next.

#### D. Performance measures and data requirements

We assume that spiking responses of the system are less frequent than periods of silence and therefore the accuracy of our reconstruction should depend of the number of spikes  $N_{\text{spikes}}$  produced by the model (or recorded data in the experimental context), independent of the actual threshold. Further, we expect that when the stimulus dimensionality  $D$  increases, the fact that there are more parameters needed to describe the kernel  $K$  means that performance will deteriorate unless we have access to more data. To address these issues we explore model neurons as in Fig. 3, but systematically vary  $D$  and  $N_{\text{spikes}}$ . Again we consider the most difficult case, with naturalistic stimuli and kernels that are random symmetric matrices. The results are shown in Fig. 4.

Our intuition is that the number of spikes should scale with the number of free parameters in the model,  $D(D+1)/2$ , so we always start with  $N_{\text{spikes}} = D(D+1)/2$ . If we normalize the reconstruction errors by this initial error, and scale  $N_{\text{spikes}}$  by the number of free parameters, the data collapse, as shown in Fig. 4b. Evidently accurate reconstructions of very large matrices requires very many spikes. This suggests that it will be important to have some more constrained models, which we now explore.

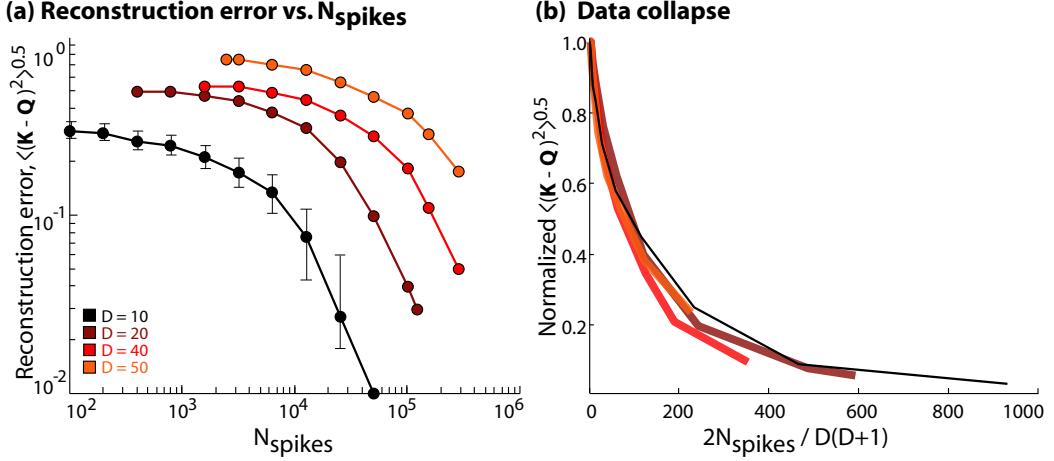


FIG. 4. **Data requirement and performance issues.** (a) Reconstruction error  $\langle(\mathbf{K} - \mathbf{Q})^2\rangle^{1/2}$  is plotted as a function of number of spikes ( $N_{\text{spikes}}$ ) for matrices corresponding to stimuli of increasing dimensionality ( $D = 10, 20, 40$  &  $50$ ). (b) The traces for different values of  $D$  collapse when the error is normalized by the (maximum) value at  $N_{\text{spikes}} = D(D+1)/2$  for each  $D$ , and plotted as a function of  $2N_{\text{spikes}}/(D(D+1))$ .

#### IV. CONSTRAINED FRAMEWORKS

The most general stimulus energy is described by  $D(D+1)/2$  parameters, and this quickly becomes large for high dimensional stimuli. In many cases it is plausible that the matrix kernel of the stimulus energy has some simpler structure, which can be used to reduce the number of parameters.

One way to simplify the description is to use a matrix that has low rank. If, for example, the rank of the matrix  $\mathbf{Q}$  in Eq. (23) is  $p \leq D$ , then we can find a set of

orthogonal vectors  $\{\mathbf{v}_i\}$  such that

$$\mathbf{Q} = \sum_{i=1}^p \mathbf{v}_i \mathbf{v}_i^T. \quad (42)$$

In terms of these vectors, the stimulus energy is just  $x = \sum_{i=1}^p (\mathbf{v}_i^T \cdot \mathbf{s})^2$ .

The low rank approximation reminds us of the simpler, Euclidean notion of dimensionality reduction discussed above. Thus, we could introduce variables  $x_i = \mathbf{v}_i^T \cdot \mathbf{s}$  for  $i = 1, 2, \dots, p$ . The response would then be approximated as depending on all of these variables,  $r(x_1, x_2, \dots, x_p)$ , as in Eq. (7). In the stimulus energy approach, all of these multiple Euclidean projections are combined into  $x = \sum_i x_i^2$ , so that have a more constrained but potentially more tractable description. When  $\mathbf{Q}$  is written as  $\mathbf{Q} = \sum_{i=1}^p \mathbf{v}_i \mathbf{v}_i^T$ , the relevant gradient of information, analogous to Eq. (37) is

$$\nabla_{\mathbf{v}_i} I = 2 \int dx P_{\mathbf{Q}}(x) [\langle \mathbf{s}(\mathbf{v}_i^T \mathbf{s}) | x, \text{spike} \rangle - \langle \mathbf{s}(\mathbf{v}_i^T \mathbf{s}) | x \rangle] \frac{d}{dx} \left[ \frac{P_{\mathbf{Q}}(x | \text{spike})}{P_{\mathbf{Q}}(x)} \right], \quad (43)$$

and we can turn this into an algorithm for updating our estimates of the  $\mathbf{v}_i$ ,

$$\mathbf{v}_i \rightarrow \mathbf{v}_i + \gamma \nabla_{\mathbf{v}_i} I. \quad (44)$$

There is a free direction for the overall normalization of the matrix  $\mathbf{Q}$  [6, 17] which makes the mutual information invariant to reparameterization of the quantities. We orthogonalize the vectors  $\mathbf{v}_i$  during the learning procedure to make sure that they don't grow out of bound.

Another way of constraining the kernel of the stimulus energy is to assume that it is smooth as we move from one stimulus dimension to the next. Smooth matrices can be expanded into weighted sums of basis functions,

$$\mathbf{Q} = \sum_{\mu=1}^M \alpha_{\mu} \mathbf{B}^{(\mu)}, \quad (45)$$

and finding the optimal matrix then is equivalent to cal-

culating the most informative  $M$ -dimensional vector of weights.

The basis can be chosen so that systematically increasing the number of basis components  $M$  allows the reconstruction of progressively finer features in  $\mathbf{Q}$ . For example, we can consider  $\{\mathbf{B}^{(\mu)}\}$  to be a set of Gaussian bumps tiling the  $D \times D$  matrix  $\mathbf{Q}$ , and whose scale (standard deviation) is inversely proportional to  $\sqrt{M}$ . For  $M \rightarrow D^2/2$  the basis matrix set becomes a complete basis, allowing every  $\mathbf{Q}$  to be exactly represented by the vector of coefficients  $\alpha$ . In any matrix basis representation, the learning rule becomes,

$$\alpha_\mu \rightarrow \alpha_\mu + \gamma \sum_{i,j=1}^M \frac{\partial I}{\partial \mathbf{Q}_{ij}} \mathbf{B}_{ij}^{(\mu)}. \quad (46)$$

This is equivalent to taking projections of our general learning rule, Eq. (38), onto the basis elements.

## V. RESULTS FOR MODEL NEURONS

### A. An auditory neuron

As a first example, we return to the model auditory neuron whose response properties from Eq's (8) and (9) were schematized in Fig. 1. Rather than studying its responses to white noise stimuli, however, we consider the responses to bird song, as shown in Fig. 5(a). We start with Eq. (8) and see that it is equivalent to a stimulus energy with kernel  $\mathbf{K}$  defined through

$$p(t) = \int dt_1 \int dt_2 s(t_1) \mathbf{K}(t - t_1, t - t_2) s(t_2), \quad (47)$$

$$\mathbf{K} = \int d\tau f_2(\tau) f_1(t - \tau - t_1) f_1(t - \tau - t_2). \quad (48)$$

We used the same filters as we showed in Fig. 1(b) and (e) to construct  $\mathbf{K}$ , which is plotted in Fig. 5(b). We can also look in a mixed time-frequency representation to generate a spectrotemporal “sensitivity,”  $\tilde{\mathbf{K}}$ , as follows:

$$\tilde{\mathbf{K}}(\omega, t) = \int d\tau \mathbf{K}\left(t + \frac{\tau}{2}, t - \frac{\tau}{2}\right) e^{+i\omega\tau}. \quad (49)$$

Eq. (49) suggests that  $\tilde{\mathbf{K}}$  is a function of  $\omega$  and  $t$ , as seen in Fig. 5(c).  $\tilde{\mathbf{K}}$  describes the selectivity of the model neuron to the stimulus, in this case the bird song. We see that this neuron responds robustly to a sound with a frequency around 1KHz with a temporal dependence dictated by the time constants of the 2 filters that make up the neuron's receptive field matrix  $\mathbf{K}$ . This description has the flavor of a spectrotemporal receptive field (STRF), but in the usual implementations of the STRF idea a spectrogram representation is imposed onto the stimulus, fixing the shapes of the elementary bins in the time-frequency plane. Here, in contrast, Fourier transforms are in principle continuous, and we don't need to

assume that the only relevant variable is stimulus power in each frequency band.

The natural stimuli we used to probe this model auditory neuron's receptive field came from recordings of zebra finch songs, modified into stimulus clips  $\mathbf{s}$ . The songs were interpolated down from their original sampling rate to retain the same discrete time steps ( $dt = 1/20000$ s) that we use in Fig. 1. An example sound pressure wave of a song stimulus is plotted in Fig. 5(a). In the same panel, we emphasize the structural complexity of the stimulus in the spectrogram of the same song. The song spectrogram shows multiple motifs, significant temporal structure and several harmonic stacks, all of which point to the strong departure from Gaussianity of  $\mathbf{s}$ .

This model neuron, as illustrated in Fig. 1, emitted spikes when the power in Eq. (48) exceeded a threshold. We set the threshold of firing so that the mean spike rate was  $20\text{s}^{-1}$ . We presented  $\sim 42$  minutes of bird song stimuli to the model neuron, collecting roughly 50,000 spikes.

We follow the gradient ascent procedure exactly as described in Eq. (38). Note that  $\mathbf{K}$  in this model is a  $300 \times 300$  matrix, and we make no further assumptions about its structure; we start with a random initial condition (Fig. 5d). The maximally informative matrix that we find is shown in Fig. 5e, and is in excellent agreement with the matrix  $\mathbf{K}$ ; we can see this in the frequency domain as well, shown in Fig. 5f. Quantitatively, the RMS reconstruction error for this inference is less than 5% of the maximum for any two uncorrelated random, symmetric matrices of the same size.

### B. Complex cell in the visual cortex

We consider the model complex cell described earlier in Eq. (19), but allow integration over just one frame of a movie, so we don't need to describe the temporal filter. We chose parameters  $k = 2\pi/3$ ,  $\sigma_1 = 1.6$  and  $\sigma_2 = 5$ , with positions measured in pixels. The stimuli were 20,000 grayscale  $30 \times 30$  pixel image patches extracted from a calibrated natural image database [26]. Spikes were generated with a probability of a spike/bin of 0.1 whenever the stimulus power  $p$  exceeded a threshold. We note that, in this model, the kernel is explicitly of rank 2, and so we followed the algorithm in Eq. (43). The results are shown in Fig. 6. As expected, the best possible reconstruction is a vector pair  $\mathbf{v}_1, \mathbf{v}_2$  that is equal to the pair  $F_1, F_2$  up to a rotation. Fig. 6 shows that this is indeed the case for the reconstructed filters  $\mathbf{v}_1, \mathbf{v}_2$ , which otherwise match the true filters well.

Suppose that the real neuron, as in our model, is described by a kernel of rank 2, but we don't know this and hence search for a kernel of higher rank. As shown in Fig. 7c, higher rank fits do not increase the information that we capture, either for random stimuli or for natural stimuli. Interestingly, however, the “extra” components of our model are not driven to zero, but appear as (re-

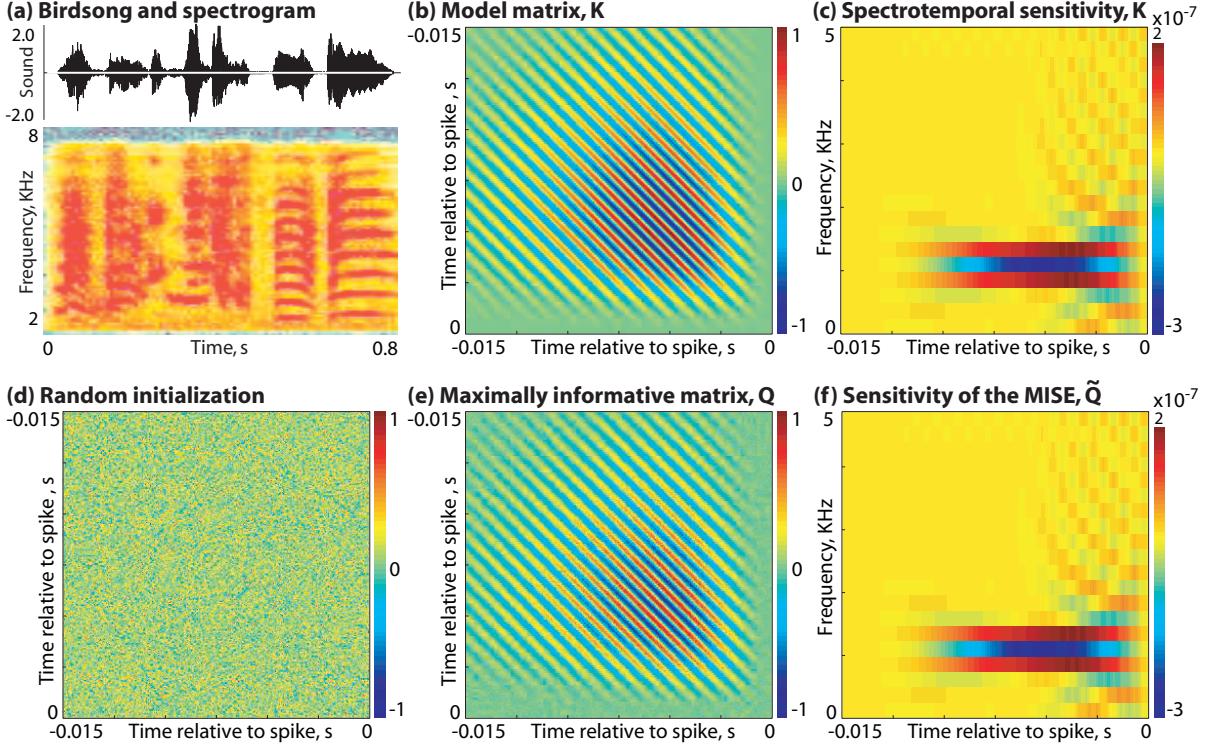


FIG. 5. **Analyzing the responses of the model auditory neuron to a bird song.** (a) The sound pressure wave of a zebra finch song used as stimulus to the model neuron is shown along with its spectrogram. The spectrogram of the song illustrates that  $\mathbf{s}$  is highly structured, full of harmonic stacks and complex spectrotemporal motifs. (b) The equivalent matrix  $\mathbf{K}$ , constructed from the two filters as described in Eq. (8) is  $300 \times 300$  in size but has a relatively simple structure. (c) Taking a Fourier transform over  $t_2$  of  $\mathbf{K}$  yields a spectrotemporal sensitivity matrix,  $\tilde{\mathbf{K}}$  with a peak at approximately 1KHz. (d) The initial guess for  $\mathbf{Q}$  is the random symmetric matrix plotted here. (e) The optimal matrix  $\mathbf{Q}$  that maximizes the mutual information between the spiking response of the model neuron and the 1D projection  $x = \mathbf{s}^T \cdot \mathbf{Q} \cdot \mathbf{s}$  matches  $\mathbf{K}$  well at the end of 100 learning steps. (f) The spectrotemporal sensitivity  $\tilde{\mathbf{Q}}$ , corresponding to the maximally informative stimulus energy has the same response preferences as  $\tilde{\mathbf{K}}$ .

dundant) linear combinations of the two true underlying vectors, so that the algorithm still finds a genuinely two dimensional, albeit over-complete, solution.

The convergence of a full rank matrix  $\mathbf{Q} = \sum_{i=1}^2 \mathbf{v}_i \mathbf{v}_i^T + \eta$ , where  $\eta$  is Gaussian random noise of  $\mathcal{O}(D)$ , to the model matrix  $\mathbf{K}$  can be determined by looking at the projections of the leading eigenvectors of the matrix  $\mathbf{Q}$  at the end of the maximization algorithm. In this complex cell example where spikes are generated from a rank 2 matrix, the two eigenvectors corresponding to the two leading eigenvalues for the fit,  $\mathbf{Q}$  should be identical to what used to be  $\mathbf{v}_1$  and  $\mathbf{v}_2$  before. The remaining eigenvalues should be driven to be 0, and this is indeed what we see in Fig. 7d.

### C. Matrix basis formalism

We illustrate this simplification by making use of the matrix basis expansion from Eq. (45) to infer a matrix  $\mathbf{K}$  that is of arbitrarily high rank. For  $\mathbf{K}$  we used a symmetrized  $250 \times 250$  pixel image of a fluid jet as shown in Fig. 8a. While this is not an example of a receptive field from biology, it illustrates the validity of our approach even when the response has an atypical and complex dependence on the stimulus. Spikes were generated by thresholding the energy  $\mathbf{s}^T \cdot \mathbf{K} \cdot \mathbf{s}$ , and the same naturalistic visual stimulus ensemble was used as before. Gaussian basis matrices shown in the inset of Fig. 8b were used to represent the quadratic kernel, reducing the number of free parameters from  $\sim 6 \times 10^4$  to  $M = 225$ . We start the gradient ascent with a large  $\gamma$  value of 1 and progressively scale it down to 0.1 near the end of the algorithm; Fig. 8b shows the information plateauing in about

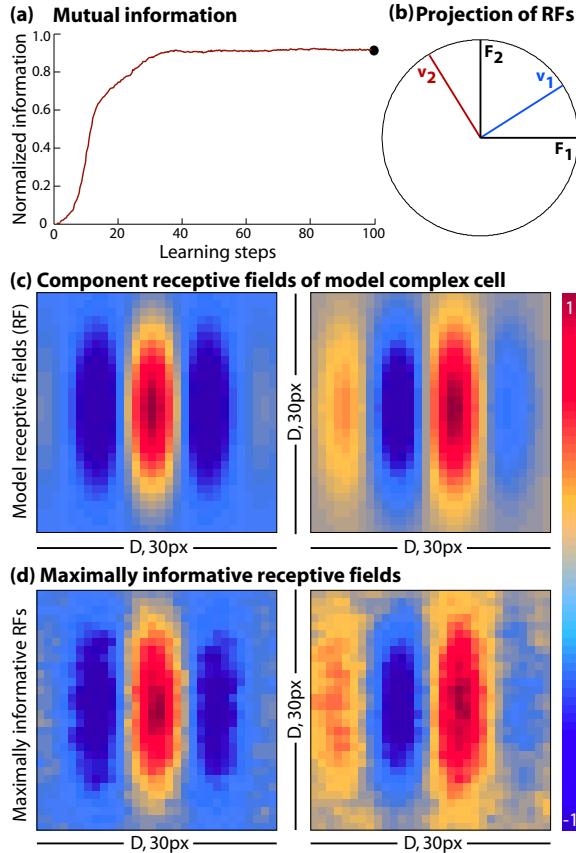


FIG. 6. **Receptive fields of a model complex cell and the reconstructed maximally informative pair.** (a) Information as a function of the number of learning steps peaks and then plateaus. The black dot is the point where the reconstructed receptive fields are shown in panel (d) below. (b) Reconstructed vectors  $v_1$ ,  $v_2$  are rotated versions of the receptive fields  $F_1$ ,  $F_2$ , but span the same linear subspace (all vectors are normalized to unit length). (c) The receptive field of the model complex cell is given by the two linear filters in Eq. (19):  $F_1$  (left) and  $F_2$  (right). (d) The reconstructed receptive fields at the 100<sup>th</sup> learning step (black dot in panel (a) above) with filters  $v_1$  (left) and  $v_2$  (right) rotated to best align with the  $F_1$  –  $F_2$  pair.

40 learning steps. The maximally informative quadratic kernel  $\mathbf{Q}$  reconstructed from these 225 basis coefficients is shown in Fig. 8c. Optimizing the 225 basis functions captures the overall structure of the kernel but this can be improved to an almost perfect reconstruction (at a pixel-by-pixel resolution) by increasing  $M$ , as shown in Fig. 8d.

#### D. Multiple matrices

As a final test of our approach, we implemented the model neuron we described earlier in Eq. (39), with stimulus dimensionality  $D = 5$ ; the two matrices,  $\mathbf{K}_1$  and

$\mathbf{K}_2$ , are plotted in Fig. 9a, along with a schematic of the spikes produced when the nonlinearity  $g(\cdot)$  is a threshold. Again we constructed stimuli from nearby pixels of natural images, so that the distribution is strongly correlated and non-Gaussian; the threshold was set so that the probability of a spike/bin to 0.3, and we generated  $\sim 10,000$  spikes. We followed the algorithm in Eq's (40) and (41) to find the maximally informative matrices  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$ . The original and the optimal matrices are plotted in Fig. 9b, where we see that for a model neuron with a randomly initialized duet of matrices, the optimal matrices agree with those of the model neuron. Mutual information in red, normalized by the maximum when  $\mathbf{K}_1 = \mathbf{Q}_1$  and  $\mathbf{K}_2 = \mathbf{Q}_2$ , and RMS reconstruction error in green (calculated as  $\langle [(\mathbf{K}_1 + \mathbf{K}_2) - (\mathbf{Q}_1 + \mathbf{Q}_2)]^2 \rangle^{1/2}$ ) are plotted as a function of the learning steps in Fig. 9c. While convergence is definitely possible, our estimates of the maximally informative matrices are noisier than in the single matrix instances, even with a relatively large amount of data, as indicated in Fig. 9d. We expect that realistic searches for multiple stimulus energies will require us to impose some simplifying structure on the underlying matrices.

## VI. DISCUSSION

While the notion that neurons respond to multiple projections of the stimulus onto orthogonal filters is powerful, it has been difficult to develop a systematic framework to infer a neuron's response properties when there are more than two filters. To get around this limitation, we propose an alternative model in which the neural response is characterized by features that are quadratic functions of the stimulus. In other words, instead of being described by multiple linear filters, the selectivity of the neuron is described by a single quadratic kernel. The choice of a quadratic form is motivated by the fact that many neural phenomena previously studied in isolation can be viewed as instances of quadratic dependences on the stimulus. We presented a method for inferring maximally informative stimulus energies based on information maximization. We make no assumptions about how the quadratic projections onto the resulting matrices map onto patterns of spiking and silence in the neuron. This approach yields unbiased estimates for receptive fields for arbitrary ensembles of stimuli, but requires optimization in a possibly rugged information landscape. The methods we have presented should help elucidate how sensitivity to high-order statistical features of natural inputs arises in sensory cortical areas.

## VII. ACKNOWLEDGMENTS

We thank AJ Doupe, RR de Ruyter van Steveninck, SR Sinha and G Tkačik for helpful discussions and for providing the natural stimulus examples used here. We

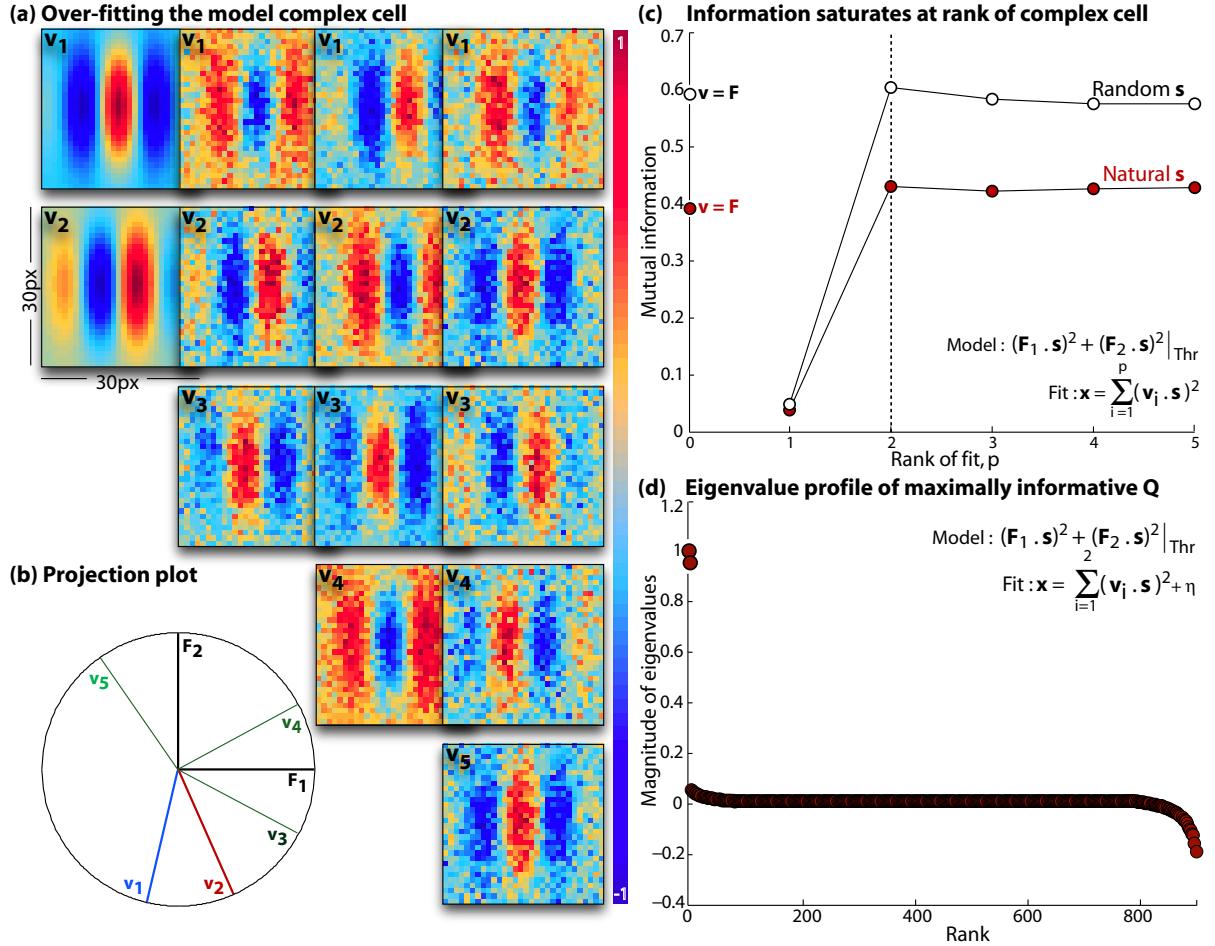


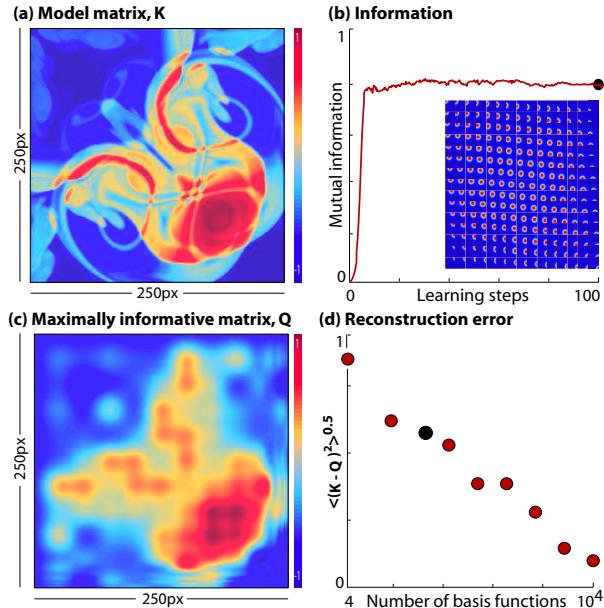
FIG. 7. **Over-fitting the model complex cell with matrices of successively increasing rank.** (a) Receptive fields reconstructed after mutual information is maximized with matrices of rank  $p = 2, 3, 4$  and  $5$  (from left to right). (b) The resulting vectors,  $v_1$  through  $v_5$ , at the end of the information maximization are no longer orthogonal but project fully into a unit circle in the  $F_1-F_2$  plane. (c) Maximum mutual information as a function of the rank of fit,  $p$ , for random stimuli (open circles) or for the stimulus matrix generated using natural scenes (filled circles), peaks at the rank equal to that of the “data” (rank  $p = 2$  for the model complex cell), and remains unchanged as the rank of  $Q$  increases. (d) Over-fitting the model matrix  $K$  with Gaussian noise does not add to the mutual information  $I_Q$  and the algorithm successfully finds a two dimensional solution. Eigenvalue profile of matrix  $Q = \sum_{i=1}^2 v_i v_i^T + \eta$  where  $\eta$  is a  $30 \times 30$  sized  $\mathcal{N}(0, 1)$  after maximizing information with respect to the complex cell. Aside from two leading eigenvalues with magnitude 1, the rest  $\rightarrow 0$ .

are also grateful to G Tkačik and O Marre for providing critical scientific input during the course of this project.

This work was supported in part by the Human Frontiers Science Program, by the Swartz Foundation, and by NSF Grant PHY-0957573.

---

- [1] HB Barlow (1972) Single units and sensation: a neuron doctrine for perception. *Perception* **1**, 371–394.
- [2] HB Barlow (1995) The neuron in perception. In *The Cognitive Neurosciences*, MS Gazzaniga, ed, pp 415–4343 (MIT Press, Cambridge).
- [3] HB Barlow (1953) Summation and inhibition in the frog’s retina. *J Physiol (Lond)* **119**, 69–88 .
- [4] JY Lettin, HR Maturana, WS McCulloch & WH Pitts (1959) What the frog’s eye tells the frog’s brain. *Proc IRE* **47**, 1940–1951.
- [5] CG Gross (2002) Genealogy of the “grandmother cell.” *Neuroscientist* **8**, 512–518.
- [6] TO Sharpee, NC Rust & W Bialek (2004) Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput* **16**: 223–50.



**FIG. 8. Matrix basis formalism for constraining the number of free parameters.** (a) A complex, structured full rank matrix  $\mathbf{K}$  is generated by symmetrizing a  $250 \times 250$  pixel image of a fluid jet, and used as the “true” kernel for our model neuron. (b) Mutual information increases with the number of learning steps, peaks at the 40<sup>th</sup> step and remains unchanged thereafter. Inset shows the collection of 225 Gaussian matrix basis functions whose peaks densely tile the space of  $\mathbf{K}$ . A trial matrix is constructed as a linear sum (with coefficients  $\{\alpha_\mu\}$ ) of the basis matrices, and information optimization is performed over  $\{\alpha_\mu\}$ . The black dot at the 100<sup>th</sup> learning step is the point where  $\mathbf{Q}$  is extracted. (c) The reconstructed matrix kernel  $\mathbf{Q}$  after maximizing mutual information using the 225 basis coefficients making up the kernel is shown here. (d) The RMS reconstruction error  $\langle(\mathbf{K} - \mathbf{Q})^2\rangle^{1/2}$  decreases as the number of basis functions  $M$  increases from 4 to  $10^4$ . With enough data perfect reconstruction is possible as  $M$  approaches the number of independent pixels in  $\mathbf{K}$ .

- [7] E de Boer & P Kuyper (1968) Triggered correlation. *IEEE Trans Biomed Eng* **15**, 169179.
- [8] E Javel, JA McGee, W Horst & GR Farley (1988) Temporal mechanisms in auditory stimulus coding. In *Auditory Function: Neurological Bases of Hearing*, GM Edelman, WE Gall & WM Cowan., eds, pp 515–558 (Wiley, New York).
- [9] AJ Hudspeth & D P Corey (1977) Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli. *Proc Natl Acad Sci (USA)* **74**, 2407–2411.
- [10] AR Palmer & IJ Russell (1986) Phase-locking in the cochlear nerve of the guinea pig and its relation to the receptor potential of inner hair cells. *Hear Res* **24**, 1–15.

- [11] DH Johnson (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J Acoust Soc Am* **68**, 1115–1122.
- [12] W Bialek & RR de Ruyter van Steveninck (2005) Features and dimensions: Motion estimation in fly vision. *arxiv.org:q-bio/0505003*.
- [13] B Aguera y Arcas & AL Fairhall (2003) What causes a neuron to spike? *Neural Computation* **15**, 1789–807.
- [14] DH Hubel & TN Wiesel (1962) Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *J Physiol Lond* **160**, 106–154.
- [15] DH Hubel & TN Wiesel (1977) Functional architecture of macaque monkey visual cortex. *Proc Roy Soc Lond Ser B* **198**, 1–59.
- [16] A Recio-Spinoso, AN Temchin, P van Dijk, YH Fan & MA Ruggero (2005) Wiener-kernel analysis of responses to noise of Chinchilla auditory-nerve fibers. *J Neurophys* **93**, 3615–34.
- [17] NC Rust, O Schwartz, JA Movshon & EP Simoncelli (2005) Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46**, 945–956.
- [18] EH Adelson & JR Bergen (1985) Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A* **2**, 284–299.
- [19] S Hassenstein & W Reichardt (1956) Systemtheoretische analyse der zeitreihenfolgen und vorzeichenauswertung bei der bewegungsperzeption des russelka fers chlorophanus. *Zeitschrift fur Naturforschung* **11**, 513524.
- [20] N Brenner, RR de Ruyter van Steveninck & W Bialek (2000) Adaptive rescaling maximizes information transmission. *Neuron* **26**, 695–702.
- [21] SP Strong, R Koberle, RR de Ruyter van Steveninck & W Bialek (1998) Entropy and information in neural spike trains *Phys Rev Lett* **80**, 197–200.
- [22] RR de Ruyter van Steveninck & W Bialek (1988) Real-time performance of a movement sensitive in the blowfly visual system: Information transfer in short spike sequences. *Proc Roy Soc Lond B* **234**, 379414.
- [23] GB Christianson, M Sahani & JF Linden (2007) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *J Neurosci* **28**, 446 455.
- [24] P Gill, J Zhang, SMN Woolley, T Fremouw & FE Theunissen (2006) Sound representation methods for spectrotemporal receptive field estimation. *J Comput Neurosci* **21**, 520.
- [25] TM Cover & JA Thomas, *Elements of Information Theory* (John Wiley, New York, 1991).
- [26] G Tkačik, P Garrigan, C Ratliff, G Milčinski, JM Klein, LH Seyfarth, P Sterling, DH Brainard & V Balasubramanian (2011) Natural images from the birthplace of the human eye. *PLoS ONE* in press.
- [27] JD Fitzgerald, LC Sincich & TO Sharpee (2011) Minimal models of multidimensional computations, *PLoS Comput Biol*. 2011 Mar; 7(3):e1001111.
- [28] I Park & JW Pillow (2011), Bayesian spike-triggered covariance, *Advances in Neural Information Processing Systems (NIPS)* 24 (accepted).

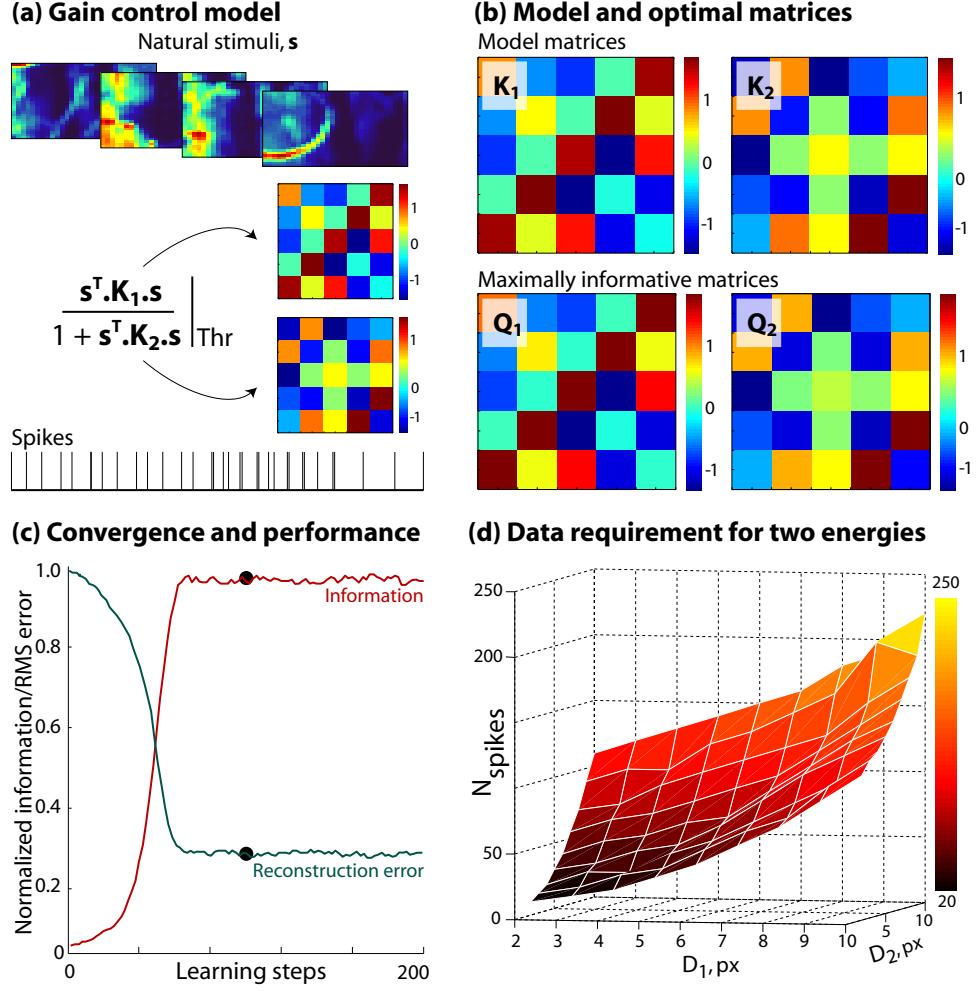


FIG. 9. **Inferring 2 MISE's simultaneously.** (a) Schematic of a model neuron with divisive gain control as described in the text. Natural images were reshaped and used as stimulus clips,  $\mathbf{s}$ . The “true” matrices  $\mathbf{K}_1$  and  $\mathbf{K}_2$  were generated as random, symmetric  $5 \times 5$  matrices. (b) The true ( $\mathbf{K}_1$  and  $\mathbf{K}_2$ , top panel) and maximally informative matrices ( $\mathbf{Q}_1$  and  $\mathbf{Q}_2$ , bottom panel) are plotted here. (c) Mutual information (red) normalized by the maximum when  $\mathbf{K}_1 = \mathbf{Q}_1$  and  $\mathbf{K}_2 = \mathbf{Q}_2$ , peaks at the 80<sup>th</sup> learning step and remains unchanged after. RMS reconstruction error (green), computed as  $\langle [(\mathbf{K}_1 + \mathbf{K}_2) - (\mathbf{Q}_1 + \mathbf{Q}_2)]^2 \rangle^{1/2}$  decays to a steady low at the same step. Using more spikes will decrease this further. (d) The data requirement for simultaneously extracting 2 matrices  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$  to a precision of 30% is shown in the shaded region.  $N_{\text{spikes}}$  here is proportional to the number of free parameters  $\sim D_1^2 + D_2^2$ , where  $D_1$  and  $D_2$  correspond to the stimulus dimensionality in the numerator and the denominator of  $x$ , respectively.